Biodiversity Data Journal 13: e153663
doi: 10.3897/BDJ.13.e153663

Research Article

Transcriptome characterisation and population genetics of *Cunninghamia konishii* Hayata - An endangered gymnosperm and implication for its conservation in Vietnam

Bich Hong Ha[‡], Mai Phuong Pham[§], Quoc Khanh Nguyen[§], Thi Tuyet Xuan Bui[|], Giap Dinh Vu[¶], Syed Noor Muhammad Shah[#], Dinh Duy Vu[§]

‡ College of Forestry Biotechnology, Vietnam National University of Forestry, Hanoi, Vietnam
§ Joint Vietnam–Russia Tropical Science and Technology Research Center, Hanoi, Vietnam
| Institute of Biology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
¶ HaUI Institute of Technology, Hanoi University of Industry (HaUI), Hanoi, Vietnam
# Department of Horticulture, Gomal University, Dera Ismail Khan, Pakistan

Corresponding author: Dinh Duy Vu (duydinhvu87@gmail.com)

Academic editor: Yuanjun Zhu

Received: 22 Mar 2025 | Accepted: 14 Jul 2025 | Published: 18 Jul 2025

Citation: Ha BH, Pham MP, Nguyen QK, Bui TTX, Vu G, Shah SNM, Vu DD (2025) Transcriptome characterisation and population genetics of *Cunninghamia konishii* Hayata – An endangered gymnosperm and implication for its conservation in Vietnam. Biodiversity Data Journal 13: e153663.
https://doi.org/10.3897/BDJ.13.e153663

Abstract

Biodiversity loss and degradation activities have a significant impact and devastating consequences on the ecosystem, eventually posing a major threat to many plant species, including *Cunninghamia konishii*. Deforestation and the growth of settlements are the main factors that affect the biodiversity hotspots in Vietnam such as Northwest, Northeast, North Central and Central Highlands regions. This has led to a decline of the species, so effective conservation strategies are urgently required. This study aimed to identify simple sequence repeat markers within expressed sequence tags from *C. konishii*, develop markers from them and assess the potential of those markers for diversity and population structure analysis of the plant. The Illumina HiSeq™ 4000 sequencing technology was applied for the transcriptomic analysis of *C. konishii* and genetic differentiation and population structure of *C. konishii* in Vietnam. In this study, the transcriptomes of *C.*

© Ha B et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

konishii were analysed using the Illumina HiSeq™ 4000 sequencing system and a total of 5,361,856,500 base pairs were generated. *De novo* assembly indicated that 58,905 unigenes were generated (average length = 736.4 bp, N50 = 1,869 bp, Q20 = 98.44% and Q30 = 95.08%). A total of 23,232 and 16,510 unigenes had significant similarities amongst Nr and Swiss-Prot, respectively. In the GO database, 12,056 (20.47%) unigenes were annotated and these genes were divided into three major categories and 50 subcategories. In the KOG analysis, 13,248 (22.49 %) unigenes were annotated and divided into 25 gene function categories. In the KEGG analysis, 8,714 (14.79%) unigenes were annotated. According to the related pathways involved, they could be classified into 56 subclasses. In this study, we have identified a total of 2,854 EST-SSRs markers. Of the 960 primer pairs, 99 were validated and reported for polymorphism. The genetic diversity within and amongst *C. konishii* populations was studied using 10 SSR markers. A sample size of 96 trees considered from three distant populations in Vietnam was analysed in this study. Our data determined *PIC* = 0.67, *Na* = 4.05, *Ne* = 2.76 and *P* = 100%. We reported moderate levels of genetic diversity with *Ho* = 0.56 and *He* = 0.58 and the fixation index value was recorded as positive for three populations (XL, HSP and PH). The bottleneck tests showed clear evidence of a bottleneck in PH population sizes. Genetic differentiation amongst populations was recorded very low (F_{ST} = 0.029), indicating gene flow (*Nm* = 8.169). This result indicates gene exchange between the populations of ancient *C. konishii* from different geographical areas and regions. The analysis of molecular variance (AMOVA) showed that high genetic variation existed within individuals (90.68%) compared to amongst populations (2.97%). A Discriminant Analysis of Principal Components (DAPC) and Bayesian clustering grouped the populations into three genetically similar clusters. Additionally, candidate genes related to essential oil biosynthesis were identified. This study provides the first EST-SSR marker-based genetic diversity and population structure analysis of *C. konishii*, offering valuable insights for breeding and conservation efforts. The findings establish a key genetic resource for future conservation strategies with the aim of preserving this endangered species.

Keywords

Cunninghamia konishii Hayata, genetic diversity, gymnosperms, Southeast Asia, EST-SSR

Introduction

Cunninghamia konishii Hayata is an essential timber species distributed mainly in four provinces of Vietnam, namely Ha Giang, Son La, Thanh Hoa and Nghe An (Pham and Vu 2022) and extends to Taiwan (China) and Laos (Farjon 2010). Morphologically, *C. konishii* is a 40–50 m tall tree with monopodial growth. The trunk diameter reaches 3-4.5 m at breast height (DBH) on full maturity. It can be found at elevations 1300 - 2800 m in Taiwan (Liang 2010), 900-2200 m in Laos (Averyanov et al. 2014) and 1000-1600 m in Vietnam (Phan et al. 2017). Both the number and size of *C. konishii* naturally occurring

populations in Vietnam are severely restricted (Phan et al. 2017) as its timber is one of the best building materials known for its fragrance (essential oil) and outstanding durability (Tran et al. 2015). Its current commercial exploitation level is very high, but cultivation is rare (Farjon 2010). There is a gradual decline in the natural population of *C. konishii* due to the destruction of the primary forests. This species is assessed as "Endangered" by the IUCN and listed on the IUCN Red List (Thomas and Yang 2013) and as endangered (EN) in the Vietnam Red Data Book 2024 (Most and Vast 2024). The complete genetics of the species remain unexplored due to the limitation of molecular marker availability. Therefore, it is of prime importance to study its distribution, ecology, natural active components, phytochemicals and other potential resources to ensure the survival of the species.

A simple sequence repeat (SSR) marker is a convenient tool in genetic diversity, population structure and evolution of plants because of its co-dominance, polymorphism and higher stability (Varshney et al. 2005). These microsatellites or SSRs consist of tandem repeat motifs of few base pairs in length, but may vary in number at a given locus and are widely distributed throughout the eukaryotic genomes (Srivastava et al. 2019). Currently, microsatellites are popular due to their multi-allelic nature, co-dominant inheritance, polymorphic, ease of detection by polymerase chain reaction and extensive genome coverage (Ueno et al. 2012). Microsatellites are also species-specific markers (Zane et al. 2002, Squirrell et al. 2003) and ideal for population structure and genetic diversity studies since they exhibit co-dominance and high level of reproducibility (Ellis and Burke 2007, Haji et al. 2014). The conventional strategy of developing SSRs through the screening of partial genomic libraries is a time-consuming, laborious and expensive process. With recent progress in bioinformatics, an alternative approach can be used other than conventional because databases can be mined for SSRs development (Kantety et al. 2002, Kumpatla and Mukhopadhyay 2005). In recent years, with the increase in quantities of sequences of genomes in the National Centre for Biotechnology Information (NCBI) database, there has also been recorded an increase in the number of expressed sequence tags (ESTs). These EST-derived SSR markers are an alternative complement to existing genomic SSRs, which can be developed in silico by data mining of EST databases. EST–SSRs have several intrinsic advantages over genomic SSRs.They are more transferable across taxonomic levels, distributed, found in coding sequences and may be related to the function of potential genes (Yu and Li 2008). EST–SSRs have been widely used in genotypes diversity (Eujayl et al. 2002, Chabane et al. 2005), genetic mapping (Luro et al. 2008), population genetic analysis (Kim and Sappington 2013) and phylogenetic studies (Stàgel et al. 2008).

The next-generation sequencing (NGS) technology, particularly high throughput in transcriptome sequencing, has offered extensive expressed sequence data for non-model plants (Egan et al. 2012, Vlk and Řepková 2017). Transcriptome sequencing is a simple and effective tool used to develop a large number of unigene-based microsatellites (SSRs) for comprehensive analysis of the genome in plants (Zhang et al. 2021). Transcriptome sequence generally represents the genetic information of gene expression and may be directly related to the gene function (Jain 2011). Recently, the

development of SSR markers, based on transcriptome sequences, has been reported in many conifer species such as *Taxus florinii*, *T. wallichiana*, *T. mairei*, *Glyptostrobus pensilis*, *Torreya grandis*, *Amentotaxus argotaenia*, *Pinus koraiensis*, *Abies alba Fokienia hodginsii Platycladus orientalis* (Liu et al. 2011, Zhu and Lou 2012, Postolache et al. 2013, Wu et al. 2017, Qin et al. 2017, Ding et al. 2017, Du et al. 2017, Zeng et al. 2018, Li et al. 2019, Ruan et al. 2019, Li et al. 2020, Götz et al. 2024). In Chinese fir (*C. lancelata*), around 100 pairs of SSR markers developed through transcriptome sequences have been reported (Huang et al. 2012, Wen et al. 2017, Duan et al. 2017, Lin et al. 2020, Götz et al. 2024), but SSR markers development through transcriptome sequences have not been reported in *C. konishii*. The genetics of environmental adaptation in *C. konishii* have not been studied. Therefore, transcriptome sequencing tool needs to be used to understand the genetic and metabolic adaptations of the species to its habitats. The study of genetic diversity within and between the population will help in the efforts to manage and conserve the species. It will contribute to the knowledge of ecology, evolution and genetics for the conservation of the species.

The Illumina HiSeq^{TM}4000 platform was used to analyse the transcriptome and develop EST-SSR markers. This will help in understanding of genetic diversity and population structure of the species, which is very important for the conservation of *C. konishii*.

Materials and Methods

Sample collection and RNA extraction

The leaves, stem and roots of five different native *C. konishii* were collected from Hoang Su Phi Nature Reserve, Ha Giang Province (northern Vietnam, 22°39'49.6" N, 104°45'35.1" E) in February 2025 (Fig. 1). The collected samples were immediately placed in liquid nitrogen and transported to the Laboratory of Molecular Biology, Vietnam-Russia Tropical Centre and stored at -80^{o}C for RNA extraction. In order to perform Illumina sequencing, total RNA was first extracted from each sample using a plant RNA kit and then processed with DNase I. These procedures followed the manufacturer's instructions (Omega Biotack, Inc.). The quality of the processed RNA samples was checked by running electrophoresis on 1% agarose gels and an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and the quantity was checked using a Nanodrop ND-2000 spectrophotometer (NanoDrop Technologies, DE, USA). Finally, equal amounts of RNA from different samples were pooled for RNA sequencing.

Illumina sequencing and de novo transcriptome assembly

RNA samples were sent to Breeding Biotechnologies Co., Ltd. for mRNA-seq library construction and sequencing using Illumina HiSeq^{TM} 4000 next-generation platform technology (Illumina, Inc., CA, USA). The quality of the raw reads was checked. The adapter contaminants with the raw reads > 5% of unknown nucleotides (N) and > 20% low-quality bases (quality value < 10) were removed to obtain high-quality clean reads.

The clean reads were de novo assembled into contigs using TRINITY (Grabherr et al. 2011). The contigs were clustered into unigenes using TGICL v. 2.1 (Pertea et al. 2003).

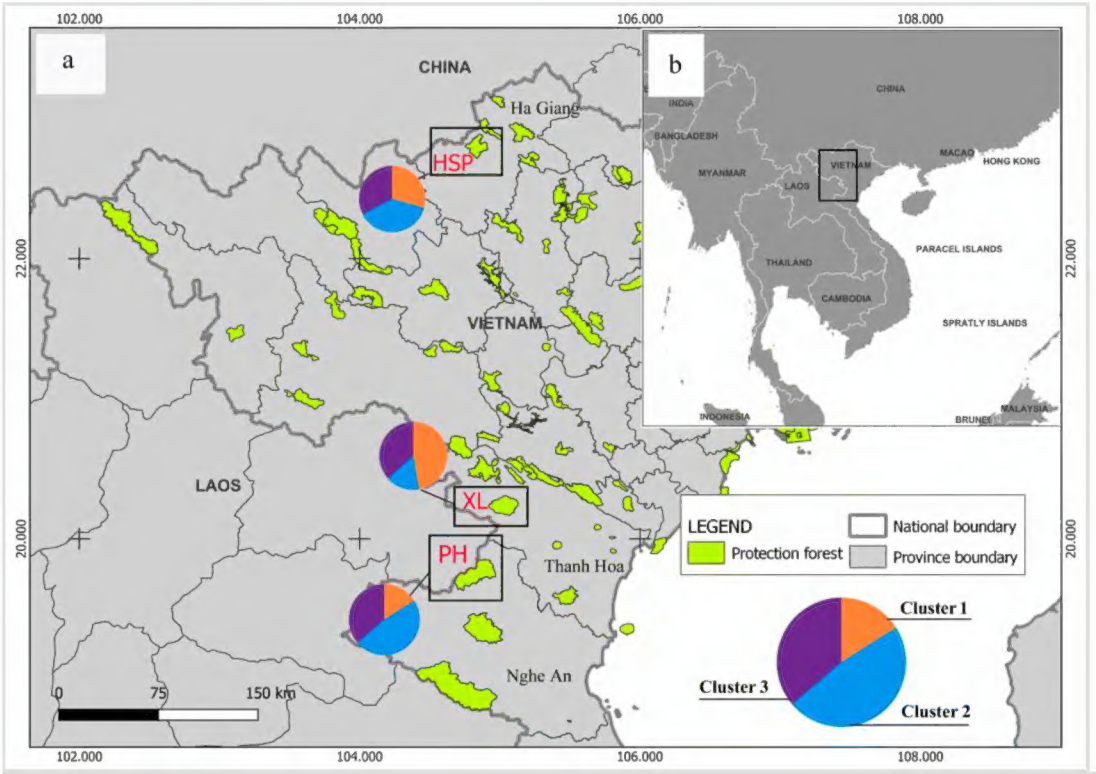

Figure 1. doi

Map of field survey locations and geographic distributions of *C. konishii* in Vietnam. Different symbols show genetic clustering into three clusters, as revealed in population structure analyses, based on microsatellite data. XL, Xuan Lien Nature Reserve; HSP, Hoang Su Phi Nature Reserve; PH, Pu Hoat Nature Reserve.

Functional annotations

To determine the predicted function, all unigene transcripts were used and compared to the NCBI non-redundant (Nr) protein database (Deng et al. 2006) and Swiss-Prot database (Apweiler 2004), Gene Ontology (GO) (Ashburner et al. 2000), eukaryotic Orthologous Groups of proteins (KOG) (Koonin et al. 2004) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa 2004) databases using BLAST (E-value of 10^{-5}) to search for homologues. Sequences were aligned with the Protein family database (Pfam) to obtain unigene annotation information using the HMMER (Finn et al. 2013). BLAST2GO software (Conesa et al. 2005) was used to determine gene ontology (GO) annotations of assembled unigenes for the categories of biological processes and molecular functions of cellular components. The BLAST software aligned the unigene sequences to the eukaryotic Orthologous Groups (KOG) clusters. KEGG pathways annotations were analysed for the metabolic pathways and the related gene functions.

Primer design

EST–SSRs with appropriate flanking regions were selected for primer design using Primer 3 software (Untergasser et al. 2012). A total of 960 primer pairs were designed and synthesised from Integrated DNA Technologies, Inc. To amplify DNA fragments

containing SSRs, the primer designing conditions were 55–65°C melting temperature, 40–60% GC content and a length of PCR product between 100 and 300 base pairs.

DNA extraction, microsatellite polymorphism identification

Amongst the 100 SSR primer pairs randomly selected for PCR identification, 10 were successfully amplified by 96 *C. konishii* individuals from three natural populations (Fig. 1, Suppl. material 6). The total genomic DNA was extracted from the plant DNA Kit according to the manufacturer's instructions (BioTeke, Beijing, China). The DNA purity and integrity were tested by a Nanodrop ND-2000 spectrophotometer (NanoDrop Technologies, DE, USA) and then diluted to 20 $ng \cdot \mu l^{-1}$. The PCR amplification reaction was performed in a 25 µl reaction volume, comprising 2.5 µl of template DNA, 12.5 µl of 2X Taq Master Mix, 1 µl of each primer and 8 µl of deionised water. The GeneAmp PCR System 9700 (Applied Biosystems, USA) was used for the PCR amplification with the following conditions: initial denaturation at 94°C for 3 minutes, 40 cycles at 94°C for 30 s, a primer-specific annealing temperature for 30 s and 70°C for 1 min, followed by a final extension at 72°C for 10 min and then holding the samples at 4°C until they were analysed. The PCR products were sized and relatively quantified between samples on a 5300 Fragment Analyzer (Agilent) with an Agilent DNF-905 dsDNA Kit (1-500 bp) (Agilent).

Data analysis

Genetic diversity

Null alleles and other genotyping errors were detected using the MICRO-CHECKER v. 2.0 software (Van Oosterhout et al. 2004), with 1000 bootstrap iterations over loci to generate the expected homozygote and heterozygote frequencies. CERCUS (Kalinowski et al. 2007) was used to estimate each locus' PIC (polymorphism information content) value. Variables for genetic diversity per loci and population, including the number of alleles (Na), number of effective alleles (*Ne*), number of private alleles (*Np*), the observed heterozygosities (*Ho*), the expected heterozygosities (*He*), inbreeding coefficient (*Fis*) and gene flow (*Nm*) were calculated using the programme GenAlEx 6.5 (Peakall and Smouse 2012). The individual inbreeding model was performed to evaluate the F_{IS} index for null allele frequency (F_{IS}IIM) using INEst (Chybicki and Burczyk 2008). Deviations from Hardy-Weinberg equilibrium (HWE) for loci within populations were tested in Cercus (Kalinowski et al. 2007), based on 1,000 permutations of alleles amongst individuals. With the use of BOTTLENECK v.1.2, each population was checked for deviations in heterozygosity from the expected values under mutation-drift equilibrium (Piry et al. 1999). Three different mutational and drift models: IAM (the infinite allele model), SMM (the stepwise mutation model) and TPM (the two-phased model of mutation) were considered. For TPM, 70% single-step mutation and 10,000 repeats were used. A one-tailed Wilcoxon signed-rank test was utilised to determine the significance of the findings.

Genetic differentiation and population structure.

The F_{ST} (Weir and Cockerham 1984) and G'_{ST} values (Hedrick 2007) were able to quantify the genetic divergence that exists between the populations (GenAlEx 6.5). Using ARLEQUIN v.3.0, the significance of the F_{ST} values in population pairs was assessed at a significance level of 0.05. The results of this testing were considered significant (Excoffier et al. 2005). The analysis of molecular variance (AMOVA) was carried out with the assistance of ARLEQUIN v.3.0, which was used to carry out the significance testing for the variance components involved. The outcomes of 10,000 different permutations served as the basis for this testing. The Mantel matrix correspondence test was used to assess the correlation between genetic distance and geographic distance amongst different geographic populations to test for a relationship between genetic distance and geographic distance (kilometres) amongst populations (Legendre and Fortin 2010). A neighbour-joining (NJ) tree was implemented, based on the Fst values using Poptree2 (Takezaki et al. 2010). To undertake an analysis of the population structure using STRUCTURE v.2.3.4, the Bayesian clustering method was put into action (Pritchard et al. 2000). Establishing the admixture model with correlated allele frequencies required ten distinct runs for each number of groups in the dataset (K). These runs were carried out with 500,000 Markov Chain Monte Carlo (MCMC) iterations and a burn-in time of 100,000 iterations. K ranged from 1 to 15. STRUCTURE HARVESTER (Earl and vonHoldt 2011) was used to detect the number of groups that best fitted the dataset, based on the K as determined by Evanno et al. (2005) to assess the ideal value of K. This was done to determine the optimal value of K. Then the duplicated findings were aligned with the help of CLUMPP v.1.1.2 (Jakobsson and Rosenberg 2007) and the allocated cluster membership bar plots were created with DISTRUCT v.1.1. (Rosenberg 2003). A Discriminant Analysis of Principal Components (DAPC) was also performed using the Adegenet package in R v.4.0.2 programme to discover clusters of individuals with a common genetic ancestor (Jombart et al. 2010). The DAPC was implemented without any previous information on the population's origin. The Bayesian information criterion (BIC) was used to establish the "ideal" cluster size and the number of clusters, denoted by the letter "K", ranged from one to twenty in the experiments. The DAPC was also responsible for performing the previous information analysis to determine how individuals should be assigned to populations. To obtain a better visual representation of the genetic clusters, the complot function of Adegenet was included. The *xval*DAPC function was used to keep the top fourteen principal components of the principal component analysis (98.5% of the variance that was conserved), as well as the seven discriminant eigenvalues.

Results

Illumina sequencing and de novo assembly

Through RNA-seq, 17,872,855 original sequences of *C. konishii* were obtained with 5,361,856,500 base pairs, Q20 (98.44%) and Q30 (95.08%). The number of 'N" bases (ambiguous base) and GC-content were 0 and 44.45%, respectively. It shows that the high-throughput sequencing platform has obtained a higher quantity and quality of *C.*

konishii sequences, which is conducive to the subsequent data assembly and meets the needs of later bioinformatics research. After the clean reads were assembled de novo, a total of 99,325 transcripts with sequence information of 105,702,940 bp and 58,905 unigenes with sequence information of 44,965,632 bp were obtained (Table 1). Analysis of the sequence length of the transcript shows that its average length was 1,064 bp and N50 was 1,869 bp. Amongst them, short sequences of 200 to 300 bp were in the majority, with 25,236 sequences accounting for 25.31% of the total; sequences with a length of 300 to 500 bp were 18,932 (19.06%) sequences; sequences of 500 to 1,000 bp in length were 18,346 (18.47%) sequences; 20,870 (21.01%) sequences were ranged from 1,000 to 2,000 bp and sequences longer than or equal to 2,000 bp accounted for 16,041 (16.15%) sequences (Suppl. material 1, Table 1). Unigene analysis and statistical results show that its average length was 763.4 bp, N50 was 1,447 bp, of which the sequences of 200 to 300 bp accounted for 21,752 (36.93%) of the total sequence; 14,338 (24.34%) sequences were between 300 and 500 bp; 9,378 (15.92%) sequences were in 500 to 1,000 bp; 7,987 (13.56%) sequences were in 1001 to 2000 bp; and the sequences of more than 2,000 bp account for 5.450 (9.25%) (Suppl. material 2, Table 1). By processing a large number of sequences obtained by high-throughput RNA-seq, the integrity of unigenes data after assembly is significantly improved and the next step of analysis and statistics were performed.

Table 1.

Length distribution of assembly transcript and unigenes for *C. konishii*.

Length range (bp)	Unigene	Transcripts
200-300	21,752 (36.93%)	25,236 (25.31%)
300-500	14,338 (24.34%)	18,932 (19.06%)
500-1000	9,378 (15.92%)	18,346 (18.47%)
1000-2000	7,987 (13.56%)	20,870 (21.01%)
> 2000	5.450 (9.25%)	16,041 (16.15%)
Total Number	58,905	99,325
Total Length	44,965,632	105,702,940
N50 Length	1,447	1,869
Mean Length	763.4	1,064

Functional annotation and classification of unigenes

A total of 58,905 Unigenes were obtained through BLAST software in 8 major databases (COG, GO, KEGG, KOG, Pfam, Swissprot, eggNOG and Nr) (Table 2). Among them, 23,232 Unigenes were successfully annotated in Nr, accounting for 39.44% of the total number of Unigenes; in KEGG, there were 8,714 successful annotations, accounting for 14.79% of the total; in SwissProt, there were 16,510 successful annotations, accounting for 28.03% of the total; 16,688 successful annotations in Pfam, accounting for 28.33% of the total; 12,056 successful annotations in GO, accounting for the total 20.47% of the

total; 13,248 successfully annotated in KOG, accounting for 22.49% of the total; 7,892 successfully annotated in COG, accounting for 13.40% of the total and in eggNOG there are 21,467 successful annotations, accounting for 36.44% of the total. The number of successfully annotated sequences in all 8 major databases was 23,878, accounting for 40.54%.

Table 2.

Functional annotation of *C. konishii* in different databases.

Annotated database	Annotated_No.	Percentage (%)	300–1000 (bp)	≥ 1000 (bp)
COG	7,892	13.40	1,591	5,357
GO	12,056	20.47	3,724	6,233
KEGG	8,714	14.79	2,411	5,244
KOG	13,248	22.49	3,818	7,438
Pfam	16,688	28.33	4,458	10,297
Swissprot	16,510	28.03	5,133	9,073
eggNOG	21,467	36.44	6,487	11,728
Nr	23,232	39.44	7,371	12,085
All	23,878	40.54	7,597	12,128

Through the Nr library's comparison, 23,232 unigenes of *C. konishii* found similar sequences in the Nr database (Suppl. material 3). The annotation matching species mainly include *Picea sitchensis* (6,658; 28.66%), *Amborella trichopoda* (2,109; 9.08%), *Nelumbo nucifera* (1,034; 4.45%), *Macleaya cordata* (571; 2.46%), *Physcomitrella patens* (552; 2.38%), *Marchantia polymorpha* (491; 2.11%), *Vitis vinifera* (375; 1.61%), *Pinus taeda* (326; 1.4%), *Elaeis guineensis* (319; 1.37%) and *Phoenix dactylifera* (301; 1.30%). A large fraction indicated similarities to genes in other species (10,445; 44.96%). From the annotated information, it can be concluded that most of the sequences of *C. konishii* can be matched in angiosperms. In general, from the distribution of sequence similarity, it can be seen that *C. konishii* has a high matching degree in the Nr database. However, due to the lack of genome and transcriptome information of *C. konishii*, no match in the database was found for some unigenes.

According to the genes successfully annotated by Nr, GO function classification annotation was performed, and the result is shown in Fig. 2. The analysis results show that 58,905 Unigenes have annotated 12,056 GO functions, accounting for 20.47% of the total number of Unigenes. Divided into three major functional categories, there are 24,768 gene sequences of cellular component functional categories, accounting for 40.35% of the total; 22,969 gene sequences of biological process functional categories, accounting for 37.42% of the total and 13,641 gene sequences of molecular functional categories, accounting for 22.22% of the total. It can be observed that the proportion of genes annotated in the functional category of a cellular component is the largest. The three functional categories can be further divided into 56 GO functional subcategories,

including 18 (cellular component), 15 (molecular function) and 20 (biological process). Among the 20 functional subcategories in biological processes: metabolic processes (6,126; 26.67%), cellular processes (5,620; 24.47%) and single-organism processes (93,799; 16.54%) have received more annotations, accounting (5,620; 24.47%), and 16.54% of this type. The percentage of annotations obtained during cell killing in biological processes was the least, having only 0.004%. In the category of cellular components, most of the annotations for cell (5,101; 20.60%), cell part (5,072; 20.50%), membrane (4.253; 17.17%), and organelle (3,659; 14.47%) have been accounted for the other organism (1), another organism part (1), and extracellular region part (13) were less annotated, accounting for 0.004%, 0.004%, and 0.05%, respectively. In the molecular function category, there are many annotations for catalytic activity (6,193) and binding (5682), each accounting for 46.4% and 41.65% of the total classification, while the annotations for translation regulator activity (1), metallochaperone activity (2) and protein tag (3) account for at least 0.007%, 0.015% and 0.022%, respectively. The above GO function annotation results show the basic situation of gene expression in *C. konishii*. It can be seen that among the 3 functional subcategories, there are more genes related to metabolic activities in biological processes, indicating that *C. konishii* has a strong metabolic ability.

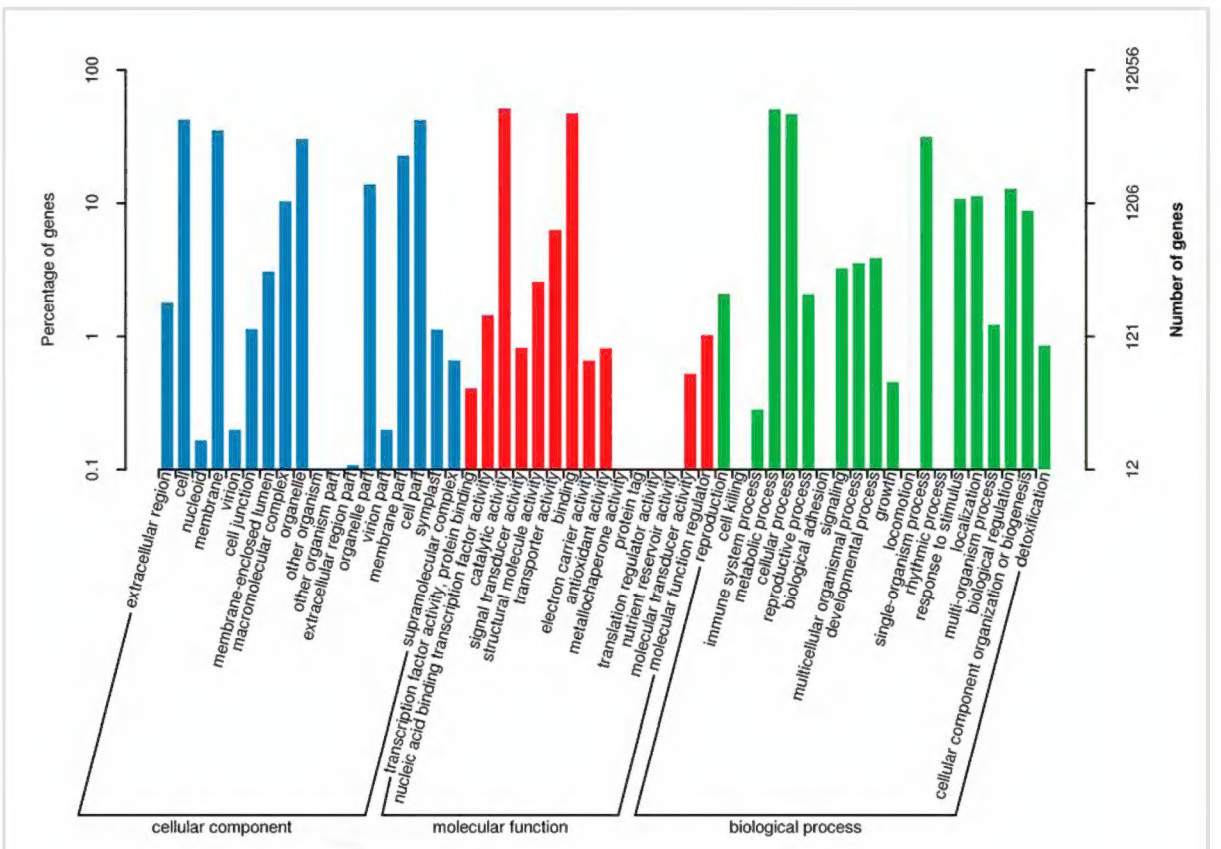

Figure 2. doi

Gene Ontology (GO) classification of unigenes in transcriptome for *C. konishii*.

The obtained Unigene is classified and annotated by the KOG protein database, and the result is shown in Fig. 3. The analysis results show that 13,248 unigenes in KOG can be matched, accounting for 22.49% of the total. The comparison results can be divided into 25 functional categories, including energy production and conversion, translation,

ribosomal structure and biogenesis and different types of gene expression, such as biosynthesis, processing and secondary metabolites biosynthesis, transport and catabolism. Amongst them, there is mainly the general function prediction only (3,939), accounting for 29.73% of the total, followed by post-translational modification, protein turnover and chaperones, accounting for 9.8% of the total; signal transduction mechanisms (8.2%); carbohydrate transport and metabolism (5.34%); and secondary metabolites biosynthesis, transport and catabolism (5.01%). The functional genes for transcription (5.23%) also account for a high proportion. The extracellular structures and cell motility have the least functional annotation information, only five and 36, which account for 0.05% and 0.27% of the total, respectively. *C. konishii* has more gene expression in terms of transcription, translation and protein transport. In addition, there is one unknown protein whose specific biological function cannot be identified, accounting for 5.43% of the total.

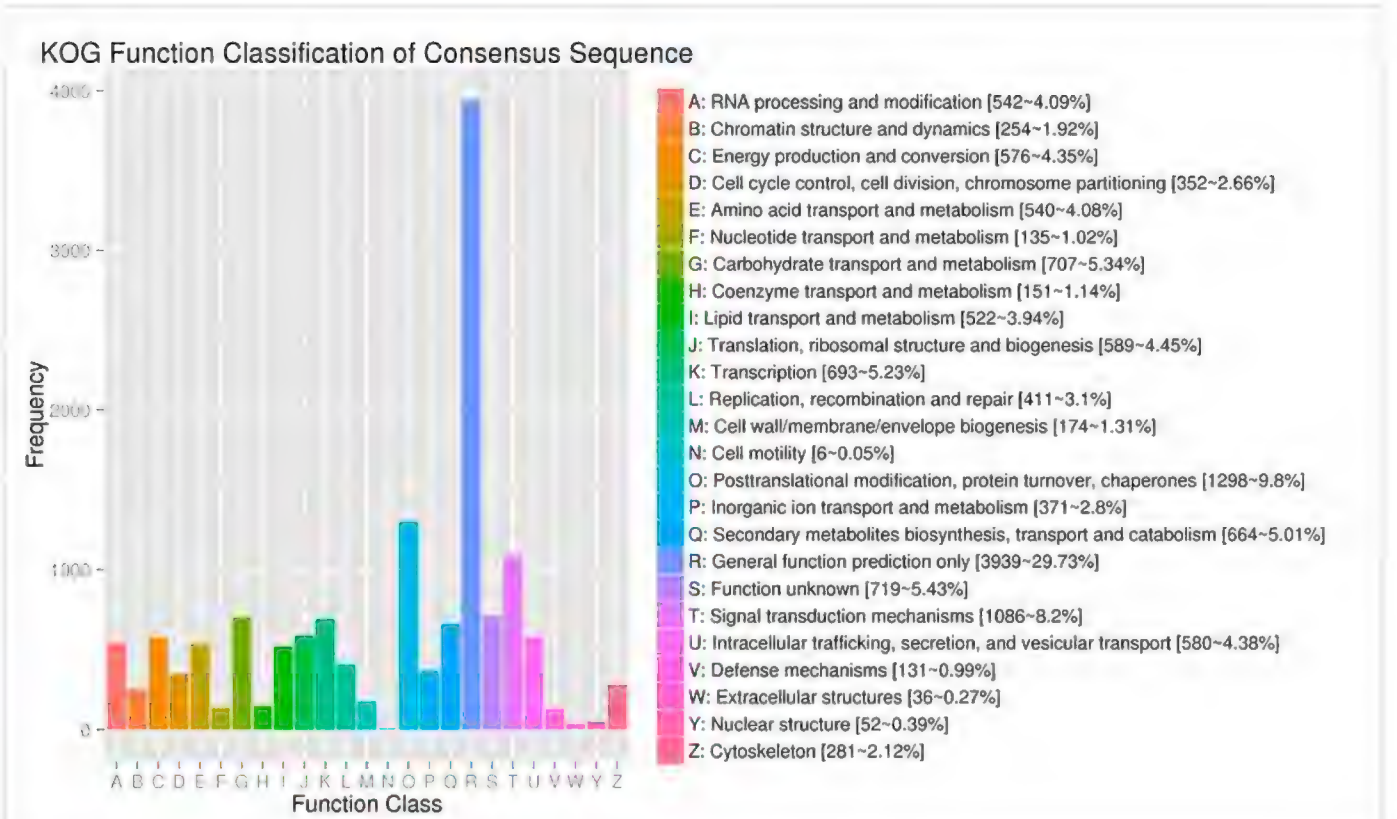
KOG Function Classification of Consensus Sequence
Frequency
4000
3000
2000
1000
0
A B C D E F G H I J K L M N O P Q R S T U V W Y Z
Function Class
A: RNA processing and modification [542~4.09%]
B: Chromatin structure and dynamics [254~1.92%]
C: Energy production and conversion [576~4.35%]
D: Cell cycle control, cell division, chromosome partitioning [352~2.66%]
E: Amino acid transport and metabolism [540~4.08%]
F: Nucleotide transport and metabolism [135~1.02%]
G: Carbohydrate transport and metabolism [707~5.34%]
H: Coenzyme transport and metabolism [151~1.14%]
I: Lipid transport and metabolism [522~3.94%]
J: Translation, ribosomal structure and biogenesis [589~4.45%]
K: Transcription [693~5.23%]
L: Replication, recombination and repair [411~3.1%]
M: Cell wall/membrane/envelope biogenesis [174~1.31%]
N: Cell motility [6~0.05%]
O: Posttranslational modification, protein turnover, chaperones [1298~9.8%]
P: Inorganic ion transport and metabolism [371~2.8%]
Q: Secondary metabolites biosynthesis, transport and catabolism [664~5.01%]
R: General function prediction only [3939~29.73%]
S: Function unknown [719~5.43%]
T: Signal transduction mechanisms [1086~8.2%]
U: Intracellular trafficking, secretion, and vesicular transport [580~4.38%]
V: Defense mechanisms [131~0.99%]
W: Extracellular structures [36~0.27%]
Y: Nuclear structure [52~0.39%]
Z: Cytoskeleton [281~2.12%]

Figure 3. doi

KOG functional annotation distribution of unigenes in transcriptome for *C. konishii*.

Comparing the obtained unigenes to the KEGG database, 8,714 unigenes were annotated, accounting for 14.79% of the total. According to the metabolic pathways involved, *C. konishii* can be classified into five major categories and 50 subcategories. The results are shown in Fig. 4. Through the specific analysis and statistics of unigenes under the relevant pathway classification in Fig. 4, it was found that the metabolic pathways account for the largest proportion of the five categories, with 2,949 unigenes, followed by pathways related to environmental information processing (1,675 unigenes), genetic information processing (231 unigenes) and cellular processes (322 unigenes); pathways related to organic systems (108) were the fewest. The five major categories were further subdivided into subcategories. Amongst them, metabolism-related pathways were divided into 30 subcategories. Carbon metabolism (250) was the majority, followed

by biosynthesis of amino acids (233) and starch and sucrose metabolism (203). The beta-alanine metabolism (52), inositol phosphate metabolism (52) and photosynthesis (52) of subcategories were the least. In addition, pathways related to genetic information processing were divided into 14 subcategories. Ribosome pathways (282) accounted for the most significant number, followed by spliceosome (218) and RNA transport (173). Base excision (52) and aminoacyl-tRNA biosynthesis (52) accounted for the smallest proportion. In the environmental information processing pathways, only two subcategories were included, with plant hormone signal transduction (177) being the majority, followed by the phosphatidylinositol signalling system (54). In the cellular processes, there were three subcategories, including endocytosis (164), peroxisome (88) and phagosome (70). Pathways related to organic systems have only one subcategory (plant-pathogen interaction), accounting for 108 unigenes. In the KEGG metabolic pathway analysis results, the category with the most annotated genes indicated that *C. konishii* has strong metabolic activities (Fig. 4).

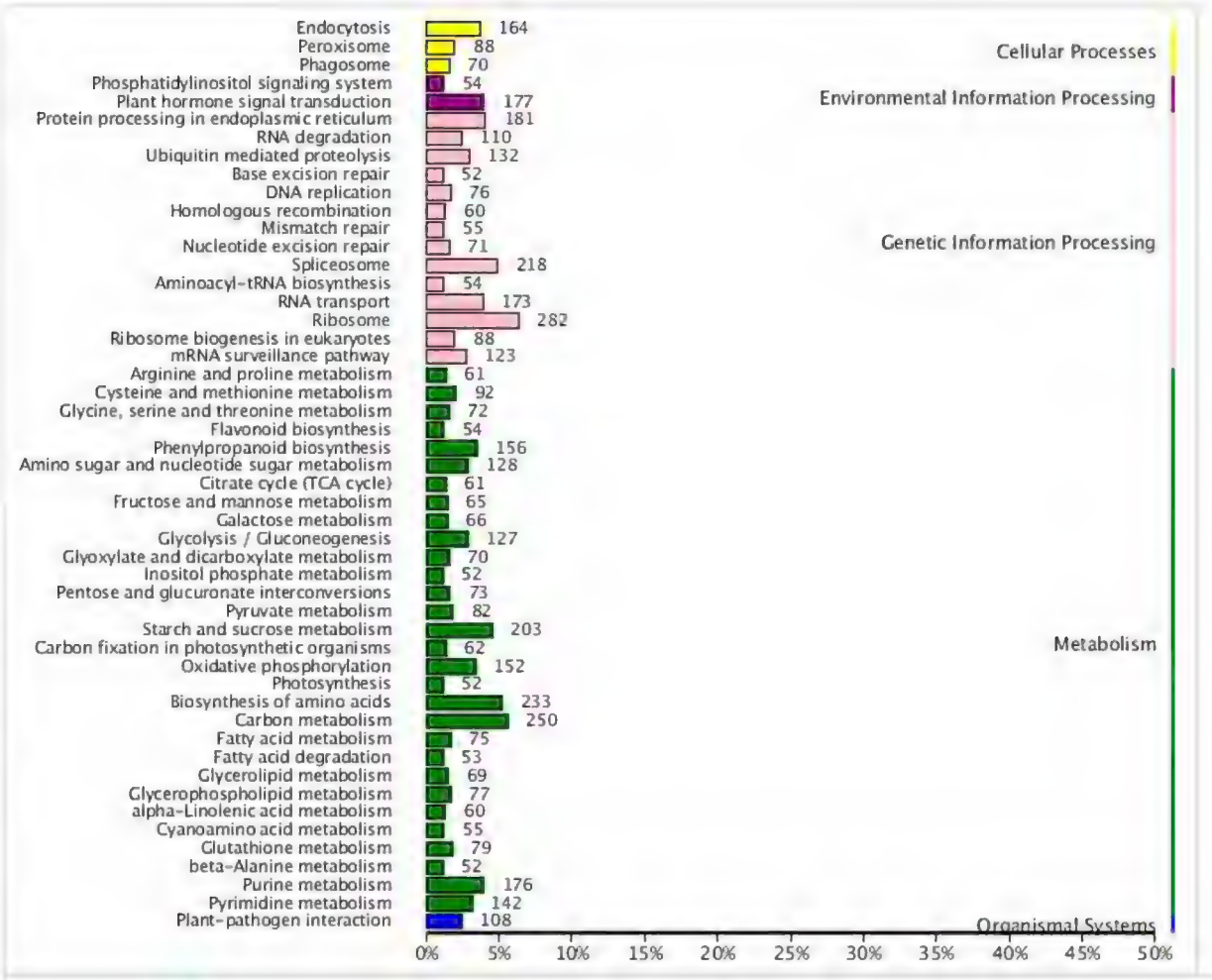

Figure 4. doi

KEGG classification of *C. konishii* unigene.

C. konishii essential oil biosynthesis candidate genes

Based on the annotation results, transcripts encoding all the known enzymes involved in essential oil biosynthesis and modification were identified in our dataset (Suppl. material 7). In most cases, multiple unigenes were annotated corresponding to the same enzyme. Such unigenes may represent different fragments of a single transcript or gene family

members. This Illumina transcriptome dataset discovered the main identified triterpene saponin biosynthetic genes from *C. konishii*. These genes include Acetyl-CoA carboxylase (EC:6.4.1.2 6.3.4.14), Enoyl-[acyl-carrier protein] reductase I (EC:1.3.1.9 1.3.1.10), Fatty acyl-ACP thioesterase A (EC:3.1.2.14) and Fatty acyl (SAD; EC:1.14.19.1). It is possible that the discovery of many genes connected to the essential oil pathway will assist us in determining what causes the high essential oil content in *C. konishii*. The discovery of many genes related to the essential oil pathway may help us to investigate the content of essential oil in *C. konishii*.

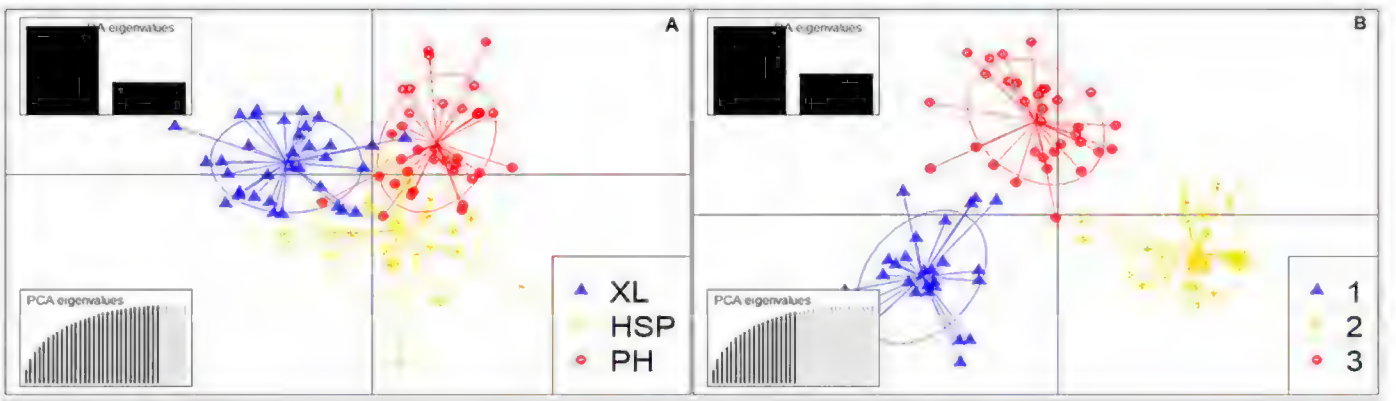

Figure 5. doi

Analysis of population structure using DAPC. Scatterplot of the DAPC with prior information (A) and Scatterplot of the DAPC without prior information (B) for *C. konishii*.

Frequency and distribution of SSRs in the unigenes

All 13,437 assembled unigenes with 27,649,667 bp were examined to discover potential EST-SSRs (Suppl. material 8). A total of 2,854 SSRs were found amongst 58,905 unigenes. Of the 2,854 SSRs, 2,413 sequences contained at least one microsatellite locus, 372 sequences contained more than one microsatellite locus and 98 SSRs involved in compound formation.

The frequency of microsatellites in the *C. konishii* unigenes was 21.24%, with an average of one SSR per 9.688 kb. Of the 2,854 potential SSRs, distribution to different repeat type classes was identified (Suppl. material 4): mononucleotide repeats and trinucleotide repeats were the most abundant, with 1,883 (65.98%) and 652 (22.85%), respectively, followed by dinucleotide repeats (287, 10.06%), tetranucleotide repeats (19, 0.67%), pentanucleotide repeats (3, 0.11%) and hexanucleotide repeats (10, 0.35%). It can be seen that the main repeat types of microsatellite sites in the *C. konishii* transcriptome are mononucleotide repeats, followed by trinucleotide repeats and dinucleotide repeats.

Regarding the number of repeat units, the distributions and frequencies of nucleotide repeats are presented in Table 3. The microsatellites that contained ten repeat units were the most prevalent (744, 26.07%), followed by those that contained five tandem repeats (439, 15.38%), eleven tandem repeats (325, 11.39%) and six tandem repeats (311, 10.90%), while the remaining tandem repeats contributed to less than 10% of the total SSR.

Table 3.

Frequency of SSRs, based on repeat types in *C. konishii* transcriptome.

Number of repeats	Repeat type						Total	Percentage (%)
	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-		
5	-	-	416	13	3	7	439	15.38
6	-	147	155	6		3	311	10.90
7	-	50	71				121	4.24
8	-	28	9				37	1.30
9	-	19					19	0.67
10	728	15	1				744	26.07
11	298	27					325	11.39
12	165	1					166	5.82
13	106						106	3.71
14	70						70	2.45
15	61						61	2.14
16	55						55	1.93
17	50						50	1.75
18	73						73	2.56
19	85						85	2.98
20	106						106	3.71
21	62						62	2.17
22	18						18	0.63
23	6						6	0.21
Total	1883	287	652	19	3	10	2854	
Percentage (%)	65.98	10.06	22.85	0.67	0.11	0.35		

Through statistical analysis of EST-SSR, based on motif types in the *C. konishii* transcriptome, the results are shown in Suppl. material 9. In the mononucleotide repeats, A/T had a large proportion (1,874, 65.66% of all SSRs), followed by C/G (9, 0.32%). In the dinucleotide, the dominant nucleotide repeats were AT/AT (134, 4.7%), followed by AG/CT (117, 4.1%) and AC/GT (36, 1.26%). In the trinucleotide, AGG/CTT was the most common
(169, 5.92%), followed by AGC/CTG (122, 4.27%), AGG/CCT (112, 3.92%), ATC/ATG (77, 2.7%), AAT/ATT (54, 1.89%), AAC/GTT (39, 1.37%), CCG/CGG (36, 1.26%), ACC/GGT (30,
1.05%), ACG/CGT (9, 0.32%) and ACT/AGT (4, 0.14%). In the tetranucleotide, AATG/ATTC was the most common (4, 0.14%), followed by AAGG/CCTT (3, 0.11%), AAAT/ATTT (2, 0.07%) and AATC/ATTG (2, 0.07%). Additionally, in the pentanucleotide and

hexanucleotide repeats, the types of repeat motifs were evenly distributed, which is related to the small number of corresponding microsatellite sites.

Genetic diversity

Of the 960 EST-SSR primers developed for *C. konishii*, 99 were able to amplify the target sequences successfully. The sequences of these 99 SSR loci were deposited into NCBI (accession numbers: MW366236 to MW366334). Using 10 different polymorphic EST-SSRs, a study was conducted on the standard genetic diversity and the population genetic structure of three different *C. konishii* populations (Suppl. material 10). Null allele frequencies were determined at four loci for *C. konishii* at the significant level of 0.05. Deviation from HWE, due to excessive heterozygosity, was significant at four loci of MP16, MP17, MP21 and MP28 in *C. konishii* (Table 4). Forty-six alleles across 96 *C. konishii* plants were produced from the ten polymorphic microsatellite markers (Table 4). Table 4 shows the values: Na = 4.03 (2.67-6.67), Ne = 2.76 (1.25-5.23), PIC = 0.67 (0.52-0.82), Ho = 0.56 (0.19-078) and He = 0.58 (0.2-0.96). The fixation index (F_{IS}) was 0.03. Positive F_{IS} values were found across all five loci, indicating an excessive number of homozygotes and suggesting inbreeding. Nonetheless, these loci exhibited high inbreeding ($p < 0.05$). It may be deduced that the inbreeding coefficient of the populations ($F_{IT} = 0.06$) suggests an excess of homozygosity in the populations (Table 5).

Table 4.

Genetic parameters in ten SSR loci for *C. konishii*.

Locus	N_a	N_e	Null allele	PIC	H_O	H_E	F_{IS}	F_{IT}	F_{ST}	P_{HWE}
MP01	5.33	2.74	No	0.76	0.62	0.62	0.01	0.04	0.03	***
MP 11	4.00	3.16	No	0.65	0.68	0.88	-0.30	-0.28	0.01	ns
MP 13	3.33	2.58	No	0.70	0.59	0.83	-0.39	-0.31	0.06	***
MP 15	4.00	2.81	No	0.64	0.64	0.96	-0.49	-0.49	0.00	***
MP 16	3.33	2.48	0.24	0.66	0.60	0.23	0.61	0.62	0.02	***
MP 17	4.00	2.05	0.28	0.62	0.51	0.35	0.30	0.31	0.01	**
MP 18	3.00	2.40	No	0.52	0.58	0.77	-0.33	-0.29	0.03	**
MP 21	6.67	5.23	0.36	0.61	0.78	0.43	0.45	0.47	0.04	***
MP 28	4.00	2.90	0.37	0.82	0.65	0.31	0.53	0.56	0.06	***
MP 30	2.67	1.25	No	0.70	0.19	0.20	-0.05	0.01	0.05	***
Mean	4.03	2.76		0.67	0.56	0.58	0.03	0.06	0.03	

Note: N_a number of alleles, *Ne* effective alleles, PIC polymorphism information content, H_O and H_E observed and expected heterozygosity, F_{IS}, fixation index, Null allele the average null allele frequency, F_{IT} coefficient of total inbreeding, F_{ST}, genetic differentiation index of Weir and Cockerham (1984), ns = not significant, $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Table 5.

Population statistics for different geographical populations of *C. konishii* investigated using ten microsatellite loci.

Populations	*P(%)*	N_a	N_e	*Np*	H_O	H_E	F_{IS}	F_{IS} IIM	P value of bottleneck		
									IAM	SMM	TPM
XL	100	4.00	2.39	0.4	0.50	0.55	0.10*	0.018	0.016	ns	ns
HSP	100	4.00	2.86	0.1	0.59	0.59	0.02	0.025	0.003	ns	0.012
PH	100	4.10	3.03	0.1	0.58	0.61	0.06	0.014	0.001	0.006	0.001
Mean	100	4.03	2.76		0.56	0.58	0.06	0.019			

*Note: N*a, mean number of alleles per locus; *N*e, mean number of effective alleles; *Np*, No. Private Alleles. P%, percentage of polymorphic loci; H_O and H_E, mean observed and expected heterozygosities, respectively; F_{IS}, inbreeding coefficient with $*p < 0.05$. $**P < 0.01$, $***P < 0.001$; ns, not significant; The individual inbreeding model was performed to evaluate the F_{IS} index for null allele frequency F_{IS}IIM; IAM, the infinite allele model; SMM, stepwise mutation model; TPM (two-phased model of mutation); XL, Xuan Lien Nature Reserve; HSP, Hoang Su Phi Nature Reserve; PH, Pu Hoat Nature Reserve.

The number of polymorphic loci varied between populations for *C. konishii*. The percentage of polymorphic loci was high (100%) in all *C. konishii* populations (Table 5). The average alleles per locus (N_a) was 4.03, ranging from 4.0 in XL and HSP to 4.1 in the PH population for *C. konishii*. The highest private alleles (N_P) level was observed in XL (N_P = 0.4). The lowest value of effective alleles (Ne) was observed in the XL population (Ne = 2.39), whereas this value was the highest in the PH population (Ne = 3.03, Ho = 0.56 and He = 0.58). The lowest values of Ho = 0.50 and He = 0.55 were detected in the XL population. The F_{IS} was 0.06 (0.02-0.1). Significantly positive F_{IS} values were observed in the XL population, showing a deficiency of heterozygotes. F_{IS}IIM = 0.019 showed the F_{IS}IIM values varied from 0.14 (PH) to 0.025 (HSP), with an average of 0.019. Both the PH and the HSP populations of the *C. konishii* species experienced a recent bottleneck in their populations (Table 5).

Genetic differentiation and genetic structure

The analysis of molecular variance (AMOVA) was implemented, based on 1023 permutations and revealed that the molecular variation attributable to differentiation amongst and within the populations for *C. konishii* (Table 6). The fixation index (F_{ST}), which can measure genetic differentiation amongst loci and averaged value was 0.03 (0.00-0.06) for *C. konishii* (Table 4), showing that 2.97% and 6.35% of the genetic variation existed between populations for *C. konishii* (Table 6). The F_{ST} value amongst populations varied from 0.014 (HSP and PH) to 0.082 (XL and PH). Significant differentiation was observed for pairwise F_{ST} values for both the *C. konishii* species ($p < 0.05$ and 0.01), except for the pair of HSP and PH (Suppl. material 11). Based on the Fst values matrix, the NJ analysis showed the two main groups (Fig. 6). The first group included the population of HSP in the northeast area. The second group included the two populations of XL and PH in the north-central area, with a bootstrap of 100%. The Mantel test based on genetic distance and geographical distance amongst populations is shown

in Suppl. material 5. Mantel analysis found significant differences in mean scores on the Nei's genetic distance and geographic distance matrix amongst populations (R^2 = 0.8492).

Table 6.

Analysis of molecular variance from natural populations for *C. konishii* species produced.

	df	Sum of squares	Variance components	Total variation (%)	Fixation indices
Amongst populations	2	17.894	0.091	2.97	F_{IS} = 0.065** F_{ST} = 0.029*** F_{IT} = 0.093***
Amongst individuals within populations	93	294.336	0.194	6.35	
Within individuals	96	266.500	2.78	90.68	
Total	191	578.729	3.061	100	

Note: df, degree of freedom; **P < 0.01, ***p < 0.001

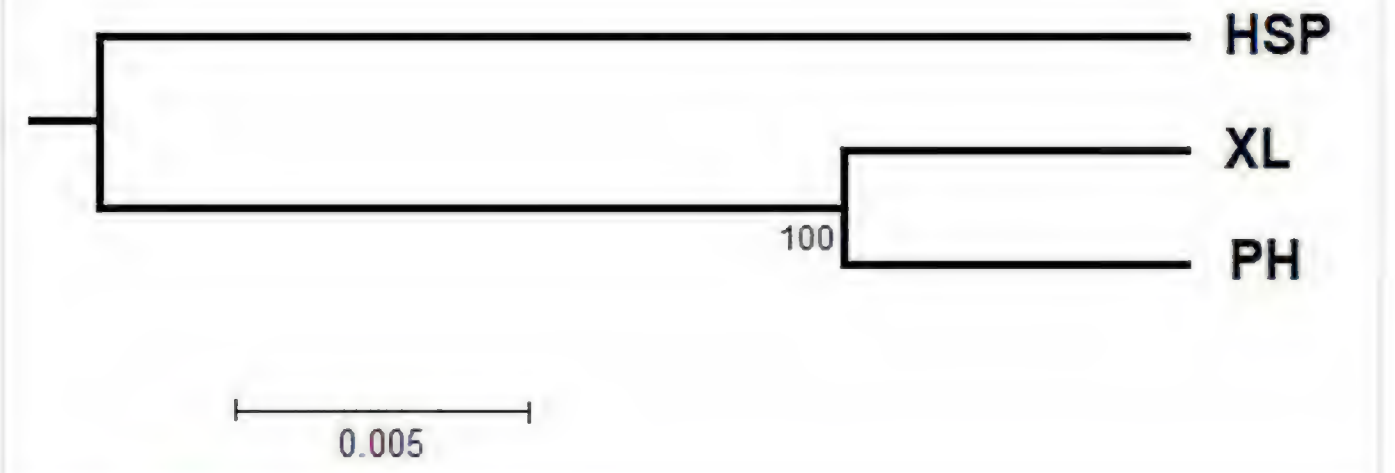

Figure 6. doi

Genetic relationships amongst populations, based on Neighbour – Joining (NJ).

Without any prior information, discriminant analysis of principal components (DAPC) also uncovered three genetic groupings for *C. konishii* (Fig. 5). Cluster 1 included 12 individuals from the XL population, seven individuals from the HSP population and 10 individuals from the PH population (Fig. 1, Suppl. material 12). Most individuals from the two populations of PH (16 individuals) and HSP (11 individuals) were assigned to cluster 2. In cluster 2, there were also seven individuals in XL. Cluster 3 included most individuals in XL and nine in HSP and PH.

The DAPC, with prior information on population origin, showed individuals within and between populations (Fig. 5). The high overlap indicated low genetic differentiation between populations. Our findings showed high overlaps between XL and HSP, with an F_{ST} value of 0.034. Similarly, low overlaps were found between HSP and PH, with an F_{ST}

value of 0.014. The Bayesian analysis of the assignation of individuals, based on the likelihoods, showed that the highest ΔK value (215.9) for 96 individuals revealed K = 3 to be an optimum number of genetic groups and showed that all individuals exhibited admixture from three groups (Figs 1, 7). The proportion of ancestry linked with each genetic group might be deduced from the colour of each individual, as measured by their percentage of the population's total colour (Fig. 7). One group (orange) was predominant in the XL populations with strong ancestry values of 47.3% and two populations (HSP and PH) had lower ancestry values of 29.4% and 16%, respectively (Suppl. material 13). In the second group (blue), two populations (PH and HSP) represented the largest group (47.7% and 37.9%, respectively), followed by the XL population (16%). Similarly, for the last group (purple), the three populations (XL, HSP and PH) were the same (36.8%, 32.7% and 36.3%, respectively). At K = 2, the *C. konishii* populations were divided into two groups. At K = 5, the central-southeast group was differentiated into five groups.

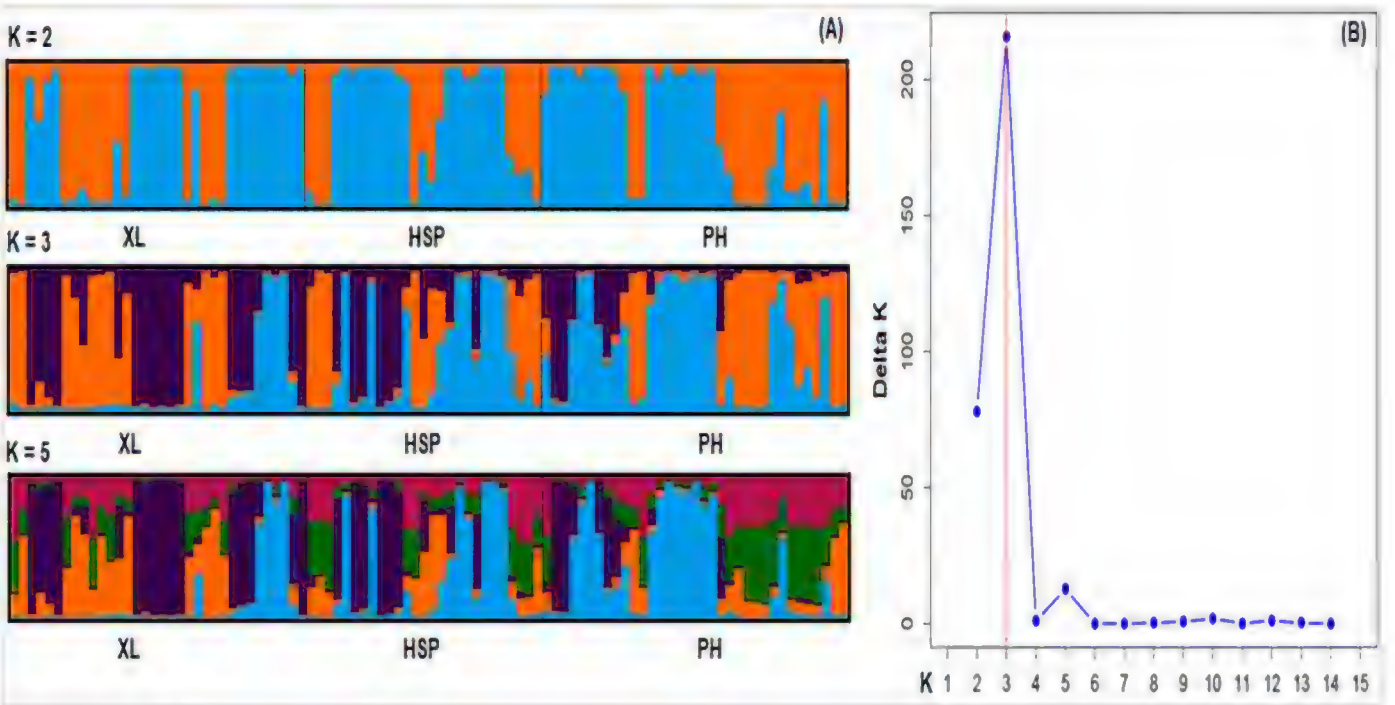

Figure 7. doi

Bar plots for three *C. konishii* populations at K = 2, K = 3 and K = 5 (A) and Mean values of distribution probability of the delta K with standard deviation from 10 runs for each value of K = 1–15 (B).

Discussion

The Illumina HiSeq4000 sequencing platform was used in this study to determine the transcriptome of *C. konishii* and bioinformatics techniques were used to analyse the microsatellite sites in the transcriptome database. The transcriptome sequencing of *C. konishii* provided a valuable resource for developing SSR markers to study genetic diversity, molecular marker-assisted breeding and the evolution of the Cupressaceae family. Fifty-eight thousand nine hundred and five (58905) transcriptomic unigenes were obtained using the Illumina HiSeq4000 platform for *C. konishii*, with a mean length of 763.4 bp and N50 of 1,447 bp. These lengths are shorter than the other conifer species, such as *Torreya grandis* (Taxaceae) with N50 = 1,503 bp and average length = 766 bp

(Zeng et al. 2018), *Cunninghamia lanceolata* (Cupressaceae) with N50 = 1,566 (Lin et al. 2020), but longer than *Amentotaxus argotaenia* (Taxaceae) with N50 = 947 bp, mean length = 721 bp (Ruan et al. 2019), indicating that the transcriptome sequencing data were well assembled for *C. konishii* in the present study. The methods used in the current study are the same. However, the deviation might be due to the sequencing depth, assembly method and natural characteristics of the species. After assembling the uni-transcripts, the GC content (44.45%) of *C. konishii* was lower than that of *Amentotaxus argotaenia* (GC = 47%) (Ruan et al. 2019), but higher than those of *Torreya grandis* (GC = 43.54%) (Zeng et al. 2018) and *Cunninghamia lanceolata* (GC = 36.04%) (Lin et al. 2020) although the methodology was the same. The GC content provides information regarding the stability of genes and genomic composition for the evolution and genetic structure of the species. These results can be related to species' adaptability to different environmental conditions. The COG, GO, KEGG, KOG, Pfam, Swiss-Prot and NR databases were searched for sequence matches in order to determine the genetic diversity of *C. konishii* and its development. Annotations of GO categories were made on the *C. konishii* unigenes. The metabolic process in the biological categories and cells in cellular components were the largest groups in this study, indicating important cellular and metabolic activities. These findings are comparable to those found in an earlier investigation of *Torreya grandis* (Zeng et al. 2018), *Amentotaxus argotaenia* (Ruan et al. 2019) and *Cunninghamia lanceolata* (Lin et al. 2020). Thirteen thousand two hundred and forty eight (13,248) unigenes were assigned to 25 KOG categories of *C. konishii*. The current KOG database results are parallel to those previously reported for 26 in KOG of *Torreya grandis* (Zeng et al. 2018) and 25 in KOG of *Amentotaxus argotaenia* (Ruan et al. 2019). For this purpose, five major categories and 50 subcategories in KEGG resources were used to study genes' biological functions and interactions. Some dominant pathways were ribosome, oxidative phosphorylation, protein processing in the endoplasmic reticulum, glycolysis/gluconeogenesis, spliceosome, RNA transport and purine metabolism. These findings suggest that the *C. konishii* unigenes discovered in the study have broad applicability and will be valuable in the future studies of the species' functional diversity.

The average length of the unigenes found in *C. konishii* was significantly longer than that of other conifers, such as *A. argotaenia* (721 bp), *Pinus pinaster* (495 bp) (Canales et al. 2013), *Platycladus orientalis* (686 bp) (Hu et al. 2016) and *P. abies* (472 bp) (Chen et al. 2012). According to Zalapa et al. (2011), having longer sequences will enhance the likelihood of successfully developing EST-SSR primers, so *C. konishii* is chosen as a prime target for microsatellite analysis in this study. Furthermore, the results indicate 58,905 unigenes in the transcriptome of *C. konishii*. There are 2,854 microsatellite loci and the frequency of occurrence is 21.24%. At the same time, there are abundant types of microsatellites in the transcriptome of *C. konishii*, with six different types of nucleotide repeats, of which the main repeat type is mononucleotide repeats, accounting for 65.98% of the total number of microsatellite loci; trinucleotide repeats, accounting for 22.85% of the total number of microsatellite loci, which are similar to *Picea* spp. (Rungis et al. 2004), *Pinus dabeshanensis* (Xiang et al. 2015), *Pseudolarix amabilis* (Geng et al. 2015), *Torreya grandis* (Zeng et al. 2018), *Amentotaxus argotaenia* (Ruan et al. 2019) and *C.*

lanceolata (Lin et al. 2020). According to Ranade et al. (2014), the AT/TA dimer motif is the most common in gymnosperms, particularly in 3' untranslated regions. Rising levels of A+T might have a role in this occurrence. As shown in Suppl. material 9, the most common trinucleotide pattern found in *C. konishii* was AAG/CTT, which is comparable to that found in *Cryptomeria japonica* (Ueno et al. 2012), *Pinus taeda* (Wegrzyn et al. 2014) and *P. halepensis* (Pinosio et al. 2014), but different from that found in *P. dabeshanensis* (Xiang et al. 2015). The AAG/CTT complex has been determined to be the methylation target in plants (Law and Jacobsen 2010).

Improvements to germplasm resources through genetic diversity have been extensively implemented in several other conifers (Zeng et al. 2018, Ruan et al. 2019, Lin et al. 2020). Generally, the level of genetic diversity within species correlates with number of loci and population, the size of its geographical range (Nguyen et al. 2020) and the extent of genetic exchange (Hellmann and Pineda-Krch 2007). According to the evaluation criteria of polymorphism (PIC > 0.5) (Botstein et al. 1980), the current study found evidence of high genetic variation, based on the EST-SSR (PIC = 0.67). Due to the relatively high PIC values, the chosen loci are both highly informative and well suited for genetic research on *C. konishii*. Several genetic measurements, including the mean number of observed alleles, observed heterozygosity, expected heterozygosity and the percentage of the polymorphic band, corroborate this finding (Lin et al. 2020). *C. konishii* exhibited a moderate degree of genetic diversity (He = 0.58; Ho = 0.56) compared to other conifers, such as *C. lanceolata* (He = 0.573; H_O = 0.526) (Lin et al. 2020), *Torreya grandis* (Ho = 0.54) (Zeng et al. 2018) and *Taxus florinii* (Ho = 0.578; He = 0.526) (Qin et al. 2017). Furthermore, our results showed that the observed number of alleles for each locus in *C. konishii* (Na = 4.03) was lower than that in *C. lanceolata* (Na =6.47) (Lin et al. 2020). The extensive genetic diversity within a species is a reflection of its evolutionary past and the ecogeographic environment in which it has thrived. It is common for species to maintain high genetic diversity when they exhibit a wide distribution pattern, a large population size, a long lifespan, a predominance of outcrossing and successional stages (Hamrick and Godt 1996, White et al. 2007). It is hypothesised that species with high genetic heterozygosity will have a greater capacity for adaptability in their environment. This hypothesis is supported by the observation that conifers have a broad distribution and exceptional longevity. However, similar results in the present study have been reported in previous studies, such as the sister species of *C. konishii*, Chinese fir - *Cunninghamia lanceolata* with Ho = 0.56 and He = 0.604 (Duan et al. 2017) and *Pinus koraiensis* with Ho = 0.299 and He = 0.311) (Li et al. 2020). Genetic heterozygosity is related to anthropogenic disturbance and can be reduced through genetic drift and increased homozygosity for common alleles due to the loss of rare alleles (Falk and Holsinger 2023). The moderate genetic heterozygosity in XL, HSP and PH suggests potential antrophogenic influence. In all studied populations, deforestation and overexploitation are key factors reducing genetic heterozygosity. Most of the conifer trees with more than 60 cm diameter were exploited by forestry enterprises in the 1980s and 1990s (Most and Vast 2024). The recent populations include trees less than 100 years old that have fragmented into small populations. The short evolutionary history and tiny population sizes of *C. konishii* may account for its moderate genetic diversity. An effective

population size and genetic variety decrease due to inbreeding may have resulted from anthropogenic disturbance and population declines. Significant heterozygosity deficits across 96 trees from three populations suggest a relatively high heterozygosity deficiency can exist within the *C. konishii* populations. This shows that inbreeding exists in small populations, although this species is outcrossed (Murawski et al. 1994). Bottleneck-determined heterozygosity deficits and bottleneck events were identified in the PH population.

In the present study, the genetic differentiation amongst *C. konishii* populations was detected by F_{ST} = 0.029, consistent with the AMOVA analysis results. The genetic differentiation of *C. konishii* was similar to that of *C. konishii* in China using AFLP markers (F_{ST} = 0.048, P = 0.016) (Chung et al. 2004). Gene flow and random mutations contribute to genetic diversity amongst populations (Schaal et al. 1998). Low population genetic differentiation reflects high gene flow. Strong gene flow (Nm > 1) determined a high number of migrants for each generation and may prevent genetic differentiation amongst populations due to genetic drift (Slatkin 1987). According to our study, a gene flow rate was comparatively high (Nm = 8.169) could mitigate the genetic drift effects by lowering the population's genetic differentiation, while increasing genetic heterozygosity. Gene flow is determined via the dispersal of pollen grains and seeds (Finkeldey and Hattemer 2007) and contributes to population genetic differences and genetic structure. Conifers are long-lived, predominantly outcrossed, insect-pollinated and late successional (Murawski et al. 1994). Their seeds are dispersed by wind and water (Ashton 1982). Insects are responsible for pollinating *C. konishii*, while the wind is responsible for spreading its seeds. The increased isolation distance will result in a reduction in the amount of pollen that is transmitted from population to population of *C. konishii*. Pollen transmission is primarily associated with insects. The current population is fragmented into smaller populations. Thus, habitat fragmentation leading to increased isolation can decrease gene exchange, which can influence its genetic structure in recent decades. The genetic structure of *C. konishii* in the present study was also detected via different clustering analyses (Structure and DAPC), in which 96 studied trees were grouped, based on their geographic origin. Therefore, the genetic structure's existence could result from gene exchange related to geographic distance and habitat fragmentation that separates populations into different genetic groups. Thus, a decrease in the dispersal of pollen and seeds might restrict gene flow for *C. konishii*.

Conclusions

The *C. konishii* transcriptome was sequenced for the first time using the Illumina 4000 sequencing technology. Many ESTs were generated and differentially expressed genes in *C. konishii* were found. A total of 2854 EST-SSRs were discovered and 99 microsatellite sites were successfully deposited into GenBank (accession numbers: MW366236 to MW366334). It is clear from the results that the naturally occurring populations of *C. konishii* have kept a reasonable amount of genetic variation over time. Several SSR markers were found, all of which will be useful in the future marker-assisted

breeding of *C. konishii*. We determined a moderate genetic variation and low genetic difference amongst the three populations of *C. konishii*. Without prior information, the structure and the DAPC identified three main genetic groups besides moderate genetic diversity in three populations of conifer species. Thus, three populations might be prioritised for in situ conservation. Simultaneously, seeds from these populations should be collected for ex-situ conservation activities. An increase in population sizes can prevent a reduction in genetic diversity via genetic drift, homozygosity for common alleles and limited gene migration across populations. This research lays the foundation for future essential oil extraction from *C. konishii* plants and conservation efforts to protect the species' genetic diversity.

Acknowledgements

This research was funded by the Basis project of the Joint Vietnam - Russia Tropical Science and Technology Research Center under grant number ST.Đ1.05/25.

Ethics and security

The authors have no competing interests to declare that are relevant to the content of this article. The authors declare that no human and/or animal material, data or cell lines were used in this study.

Author contributions

DDV, MPP and BHH: conceptualisation and design of this study; DGV, QKN, BHH, TTXB, DDV and MPP: a collection of samples; DGV, BHH, MPP and DDV: implementation of experiments; DDV, MPP, SNMS and BHH: data analysis and writing - the original draft of the manuscript; BHH, SNMS and DDV: writing - review and editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Apweiler R (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32 (90001). https://doi.org/10.1093/nar/gkh131
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for

the unification of biology. The Gene Ontology Consortium. Nature genetics 25 (1): 25-9. https://doi.org/10.1038/75556

- Ashton PS (1982) Dipterocarpaceae. In: Steenis, Ed., Malesiana Series I. 9. National Herbarium of the Netherlands, Netherlands., 237-552 pp.
- Averyanov L, Nguyen TH, Sinh KN, Pham TV, Lamxay V, Bounphanmy S, Lorphengsy S, Phan LK, Lanorsavanh S, Chantthavongsa K (2014) Gymnosperms of Laos. Nordic Journal of Botany 32 (6): 765-805. https://doi.org/10.1111/njb.00498
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics 32 (3): 314-31.
- Canales J, Bautista R, Label P, Gómez-Maldonado J, Lesur I, Fernández-Pozo N, Rueda-López M, Guerrero-Fernández D, Castro-Rodríguez V, Benzekri H, Cañas R, Guevara M, Rodrigues A, Seoane P, Teyssier C, Morel A, Ehrenmann F, Le Provost G, Lalanne C, Noirot C, Klopp C, Reymond I, García-Gutiérrez A, Trontin J, Lelu-Walter M, Miguel C, Cervera MT, Cantón F, Plomion C, Harvengt L, Avila C, Gonzalo Claros M, Cánovas F (2013) De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. Plant Biotechnology Journal 12 (3): 286-299. https://doi.org/10.1111/pbi.12136
- Chabane K, Ablett GA, Cordeiro GM, Valkoun J, Henry RJ (2005) EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Genetic Resources and Crop Evolution 52 (7): 903-909. https://doi.org/10.1007/s10722-003-6112-7
- Chen J, Källman T, Ma X, Gyllenstrand N, Zaina G, Morgante M, Bousquet J, Eckert A, Wegrzyn J, Neale D, Lagercrantz U, Lascoux M (2012) Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in norway spruce *Picea abies*). Genetics 191 (3): 865-881. https://doi.org/10.1534/genetics.112.140749
- Chung JD, Lin TP, Tan YC, Lin MY, Hwang SY (2004) Genetic diversity and biogeography of *Cunninghamia konishii* (Cupressaceae), an island species in Taiwan: a comparison with *Cunninghamia lanceolata*, a mainland species in China. Molecular Phylogenetics and Evolution 33 (3): 791-801. https://doi.org/10.1016/j.ympev.2004.08.011
- Chybicki I, Burczyk J (2008) Simultaneous estimation of null alleles and inbreeding coefficients. Journal of Heredity 100 (1): 106-113. https://doi.org/10.1093/jhered/esn088
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (18): 3674-3676. https://doi.org/10.1093/bioinformatics/bti610
- Deng YY, Li JQ, Wu SF, Zhu YP, Cai YW, He FC (2006) Integrated nr database in protein annotation system and its localization. Computer Engineering 32 (5): 71-74.
- Ding M, Meng K, Fan Q, Tan W, Liao W, Chen S (2017) Development and validation of EST-SSR markers for *Fokienia hodginsii* (Cupressaceae). Applications in Plant Sciences 5 (3). https://doi.org/10.3732/apps.1600152
- Duan H, Cao S, Zheng H, Hu D, Lin J, Cui B, Lin H, Hu R, Wu B, Sun Y, Li Y (2017) Genetic characterization of chinese fir from six provinces in southern China and construction of a core collection. Scientific reports 7 (1): 13814. https://doi.org/10.1038/s41598-017-13219-0
- Du J, Zhang Z, Zhang H, Junhong T (2017) EST–SSR marker development and transcriptome sequencing analysis of different tissues of Korean pine (*Pinus koraiensis*

Sieb. et Zucc.). Biotechnology & Biotechnological Equipment1-11. https://doi.org/10.1080/13102818.2017.1331755

- Earl D, vonHoldt B (2011) Structure harvester: a website and program for visualizing structure output and implementing the evanno method. Conservation Genetics Resources 4 (2): 359-361. https://doi.org/10.1007/s12686-011-9548-7
- Egan A, Schlueter J, Spooner D (2012) Applications of next-generation sequencing in plant biology. American Journal of Botany 99 (2): 175-185. https://doi.org/10.3732/ajb.1200020
- Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99 (2): 125-132. https://doi.org/10.1038/sj.hdy.6801001
- Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theoretical and Applied Genetics 104 (2): 399-407. https://doi.org/10.1007/s001220100738
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14 (8): 2611-2620. https://doi.org/10.1111/j.1365-294x.2005.02553.x
- Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics 1 https://doi.org/10.1177/117693430500100003
- Falk DA, Holsinger KE (2023) Genetics and conservation of rare plants. Oxford University Press https://doi.org/10.1093/oso/9780195064292.001.0001
- Farjon A, (2010) A Handbook of the world's conifers. Brill Academic Publishers, Leiden, The Netherlands https://doi.org/10.1163/9789047430629
- Finkeldey R, Hattemer HH (2007) Tropical forest genetics. Springer-Verlag, Berlin, Heidelber. https://doi.org/10.1007/978-3-540-37398-8
- Finn R, Bateman A, Clements J, Coggill P, Eberhardt R, Eddy S, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2013) Pfam: the protein families database. Nucleic Acids Research 42 https://doi.org/10.1093/nar/gkt1223
- Geng Q, Liu J, Sun L, Liu H, Ou-Yang Y, Cai Y, Tang X, Zhang H, Wang Z, An S (2015) Development and characterization of polymorphic microsatellite markers (SSRs) for an endemic plant, *Pseudolarix amabilis* (Nelson) Rehd. (Pinaceae). Molecules 20 (2): 2685-2692. https://doi.org/10.3390/molecules20022685
- Götz J, Leinemann L, Gailing O, Hardtke A, Caré O (2024) Development of a highly polymorphic chloroplast SSR set in *Abies grandis* with transferability to other conifer species-A promising toolkit for gene flow investigations. Ecology and Evolution 14 (6). https://doi.org/10.1002/ece3.11593
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29 (7): 644-652. https://doi.org/10.1038/nbt.1883
- Haji RFA, Bhargava M, Akhoon B, Kumar A, Brindavanam N, Verma V (2014) Correlation and functional differentiation between different markers to study the genetic diversity analysis in medicinally important plant *Plumbago zeylanica*. Industrial Crops and Products 55: 75-82. https://doi.org/10.1016/j.indcrop.2014.01.045
- Hamrick JL, Godt M (1996) Conservation genetics of endemic plant species. Conservation Genetics281-304. https://doi.org/10.1007/978-1-4757-2504-9_9

- Hedrick P (2007) A standarized genetic differentiation measure. Evolution 59 (8): 1633-1638. https://doi.org/10.1111/j.0014-3820.2005.tb01814.x
- Hellmann J, Pineda-Krch M (2007) Constraints and reinforcement on adaptation under climate change: Selection of genetically correlated traits. Biological Conservation 137 (4): 599-609. https://doi.org/10.1016/j.biocon.2007.03.018
- Huang H, Xu L, Tong Z, Lin E, Liu Q, Cheng L, Zhu M (2012) De novo characterization of the Chinese fir (*Cunninghamia lanceolata*) transcriptome and analysis of candidate genes involved in cellulose and lignin biosynthesis. BMC Genomics 13 (1). https://doi.org/10.1186/1471-2164-13-648
- Hu X, Liu H, Jin Y, Sun Y, Li Y, Zhao W, El-Kassaby Y, Wang X, Mao J (2016) De novo transcriptome assembly and characterization for the widespread and stress-tolerant conifer *Platycladus orientalis*. PLOS One 11 (2): e0148985. https://doi.org/10.1371/journal.pone.0148985
- Jain M (2011) Next-generation sequencing technologies for gene expression profiling in plants. Briefings in Functional Genomics 11 (1): 63-70. https://doi.org/10.1093/bfgp/elr038
- Jakobsson M, Rosenberg N (2007) Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23 (14): 1801-1806. https://doi.org/10.1093/bioinformatics/btm233
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11 (1). https://doi.org/10.1186/1471-2156-11-94
- Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16 (5): 1099-1106. https://doi.org/10.1111/j.1365-294x.2007.03089.x
- Kanehisa M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research 32 (90001). https://doi.org/10.1093/nar/gkh063
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant molecular biology 48 (5-6): 501-10. https://doi.org/10.1023/a:1014875206165
- Kim KS, Sappington T (2013) Microsatellite data analysis for population genetics. Methods in Molecular Biology271-295. https://doi.org/10.1007/978-1-62703-389-3_19
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5 (2). https://doi.org/10.1186/gb-2004-5-2-r7
- Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48 (6): 985-998. https://doi.org/10.1139/g05-060
- Law J, Jacobsen S (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics 11 (3): 204-220. https://doi.org/10.1038/nrg2719
- Legendre P, Fortin MJ (2010) Comparison of the mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology Resources 10 (5): 831-844. https://doi.org/10.1111/j.1755-0998.2010.02866.x

- Liang WY (2010) The cutting propagation technique and afforestation experiment of *Cunninghamia konishii*. Subtropical Agriculture Research 6 (4): 217-221.
- Lin E, Zhuang H, Yu J, Liu X, Huang H, Zhu M, Tong Z (2020) Genome survey of Chinese fir (*Cunninghamia lanceolata*): Identification of genomic SSRs and demonstration of their utility in genetic diversity analysis. Scientific Reports 10 (1). https://doi.org/10.1038/s41598-020-61611-0
- Liu J, Gao L, Li D, Zhang D, Möller M (2011) Cross-species amplification and development of new microsatellite loci for *Taxus wallichiana* (Taxaceae). American Journal of Botany 98 (4). https://doi.org/10.3732/ajb.1000445
- Li X, Lin X, Ruhsam M, Chen L, Wu X, Wang M, Thomas P, Wen Y (2019) Development of microsatellite markers for the critically endangered conifer *Glyptostrobus pensilis* (Cupressaceae) using transcriptome data. Silvae Genetica 68 (1): 41-44. https://doi.org/10.2478/sg-2019-0007
- Li X, Liu X, Wei J, Li Y, Tigabu M, Zhao X (2020) Development and transferability of EST-SSR markers for *Pinus koraiensis* from cold-stressed transcriptome through illumina sequencing. Genes 11 (5). https://doi.org/10.3390/genes11050500
- Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, Talon M, Ollitrault P, Morillon R (2008) Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. BMC Genomics 9 (1). https://doi.org/10.1186/1471-2164-9-287
- Most, Vast (2024) Red data book of Vietnam. Natural Sciences and Technology Publishing House [In Vietnamese].
- Murawski D, Dayanandan B, Bawa K (1994) Outcrossing rates of two endemic *Shorea* species from Sri Lankan tropical rain forests. Biotropica 26 (1). https://doi.org/10.2307/2389107
- Nguyen TM, Vu DD, Nguyen DM, Dang HP, Phan LK, Bui PX (2020) Microsatellite analysis reveals genetic diversity of the endangered species *Dipterocarpus dyeri*. Journal of Forest Research 25 (3): 198-201. https://doi.org/10.1080/13416979.2020.1747149
- Peakall R, Smouse P (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28 (19): 2537-2539. https://doi.org/10.1093/bioinformatics/bts460
- Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a softwaresystem for fast clustering of large EST datasets. Bioinformatics 19 (5): 651-652. https://doi.org/10.1093/bioinformatics/btg034
- Pham MP, Vu DD (2022) Genetic diversity and populations structure of the endangered species *Cunninghamia konishii* Hayata in Vietnam using SSR molecular markers. Research Journal of Biotechnology 17 (4): 76-83. https://doi.org/10.25303/1704rjbt7683
- Phan KL, Pham VT, Phan KL, Regalado J, Averyanov LV, Maslin B (2017) Native conifers of Vietnam - a review. Pakistan Journal of Botany 49 (5): 2037-2068.
- Pinosio S, González-Martínez SC, Bagnoli F, Cattonaro F, Grivet D, Marroni F, Lorenzo Z, Pausas JG, Verdú M, Vendramin GG (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. Molecular Ecology Resources 14 (4): 846-856. https://doi.org/10.1111/1755-0998.12232

- Piry S, Luikart G, Cornuet J (1999) Computer note. Botteneck: a computer program for detecting recent reductions in the effective size using allele frequency data. Journal of Heredity 90 (4): 502-503. https://doi.org/10.1093/jhered/90.4.502
- Postolache D, Leonarduzzi C, Piotti A, Spanu I, Roig A, Fady B, Roschanski A, Liepelt S, Vendramin GG (2013) Transcriptome versus genomic microsatellite markers: highly informative multiplexes for genotyping *Abies alba* Mill. and Congeneric Species. Plant Molecular Biology Reporter 32 (3): 750-760. https://doi.org/10.1007/s11105-013-0688-7
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2): 945-959. https://doi.org/10.1093/genetics/155.2.945
- Qin H, Yang G, Provan J, Liu J, Gao L (2017) Using Mi ddRAD-seq data to develop polymorphic microsatellite markers for an endangered yew species. Plant Diversity 39 (5): 294-299. https://doi.org/10.1016/j.pld.2017.05.008
- Ranade SS, Lin Y, Zuccolo A, Van de Peer Y, García-Gil MdR (2014) Comparative in silicoanalysis of EST-SSRs in angiosperm and gymnosperm tree genera. BMC Plant Biology 14 (1). https://doi.org/10.1186/s12870-014-0220-8
- Rosenberg N (2003) Distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4 (1): 137-138. https://doi.org/10.1046/j.1471-8286.2003.00566.x
- Ruan X, Wang Z, Wang T, Su Y (2019) Characterization and application of EST-SSR markers developed from the transcriptome of *Amentotaxus argotaenia* (Taxaceae), a relict vulnerable conifer. Frontiers in Genetics 10: 1014. https://doi.org/10.3389/fgene.2019.01014
- Rungis D, Bérubé Y, Zhang J, Ralph S, Ritland C, Ellis B, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theoretical and Applied Genetics 109 (6): 1283-1294. https://doi.org/10.1007/s00122-004-1742-5
- Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 236 (4803): 787-792. https://doi.org/10.1126/science.3576198
- Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M, Powell W (2003) How much effort is required to isolate nuclear microsatellites from plants? Molecular Ecology 12 (6): 1339-1348. https://doi.org/10.1046/j.1365-294x.2003.01825.x
- Srivastava S, Avvaru AK, Sowpati DT, Mishra R (2019) Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics 20 (1). https://doi.org/10.1186/s12864-019-5516-5
- Stàgel A, Portis E, Toppino L, Rotino GL, Lanteri S (2008) Gene-based microsatellite development for mapping and phylogeny studies in eggplant. BMC Genomics 9 (1). https://doi.org/10.1186/1471-2164-9-357
- Takezaki N, Nei M, Tamura K (2010) POPTREE2: software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. Mol. Evol. 27: 747-752. https://doi.org/10.1093/molbev/msp312
- Thomas P, Yang Y (2013) *Cunninghamia konishii*. 2013: e.T31258A2802692 https://doi.org/10.2305/IUCN.UK.2013-1.RLTS.T31258A2802692.en
- Tran HT, Ophélie B, Tran MH, Do TM, Phan KL, Nguyen TT, Félix T, Joseph C, Ange B (2015) Chemical composition of the essential oil from *Cunninghamia konishii* Hayata growing wild in Vietnam. American Journal of Essential Oils and Natural Products 2 (3): 1-5.

- Ueno S, Moriguchi Y, Uchiyama K, Ujino-Ihara T, Futamura N, Sakurai T, Shinohara K, Tsumura Y (2012) A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, *Cryptomeria japonica*. BMC Genomics 13 (1). https://doi.org/10.1186/1471-2164-13-136
- Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth B, Remm M, Rozen S (2012) Primer3-new capabilities and interfaces. Nucleic Acids Research 40 (15). https://doi.org/10.1093/nar/gks596
- Van Oosterhout C, Hutchinson W, Wills D, Shipley P (2004) microchecker: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4 (3): 535-538. https://doi.org/10.1111/j.1471-8286.2004.00684.x
- Varshney R, Graner A, Sorrells M (2005) Genic microsatellite markers in plants: features and applications. Trends in Biotechnology 23 (1): 48-55. https://doi.org/10.1016/j.tibtech.2004.11.005
- Vlk D, Řepková J (2017) Application of next-generation sequencing in plant breeding. Czech Journal of Genetics and Plant Breeding 53 (3): 89-96. https://doi.org/10.17221/192/2016-cjgpb
- Wegrzyn JL, Liechty JD, Stevens KA, Wu L, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, de Jong PJ, Mockaitis K, Main D, Langley CH, Neale DB (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196 (3): 891-909. https://doi.org/10.1534/genetics.113.159996
- Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. Evolution 38 (6). https://doi.org/10.2307/2408641
- Wen Y, Ueno S, Han W (2017) Development and characterization of 28 polymorphic EST-SSR markers for *Cunninghamia lanceolata* (Taxodiaceae) based on transcriptome sequences. Silvae Genetica 62: 137-141. https://doi.org/10.1515/sg-2013-0018
- White TL, Adams WT, Neale DB (2007) Forest genetics. Boston, MA: CABI Publishing. https://doi.org/10.1079/9781845932855.0000
- Wu X, Wen Y, Ueno S, Tsumura Y (2017) Development and characterization of EST-SSR markers for *Taxus mairei* (Taxaceae) and their transferability across species. Silvae Genetica 65 (1): 67-70. https://doi.org/10.1515/sg-2016-0009
- Xiang X, Zhang Z, Wang Z, Zhang X, Wu G (2015) Transcriptome sequencing and development of EST-SSR markers in *Pinus dabeshanensis*, an endangered conifer endemic to China. Molecular Breeding 35 (8). https://doi.org/10.1007/s11032-015-0351-0
- Yu H, Li Q (2008) Exploiting EST databases for the development and characterization of EST-SSRs in the pacific oyster (*Crassostrea gigas*). Journal of Heredity 99 (2): 208-214. https://doi.org/10.1093/jhered/esm124
- Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2011) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American journal of botany 99 (2): 193-208. https://doi.org/10.3732/ajb.1100394
- Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Molecular Ecology 11 (1): 1-16. https://doi.org/10.1046/j.0962-1083.2001.01418.x
- Zeng J, Chen J, Kou Y, Wang Y (2018) Application of EST-SSR markers developed from the transcriptome of *Torreya grandis* (Taxaceae), a threatened nut-yielding conifer tree. PeerJ 6 https://doi.org/10.7717/peerj.5606

- Zhang C, Wu Z, Jiang X, Li W, Lu Y, Wang K (2021) De novo transcriptomic analysis and identification of EST-SSR markers in *Stephanandra incisa*. Scientific Reports 11 (1). https://doi.org/10.1038/s41598-020-80329-7
- Zhu L, Lou A (2012) Development and characterization of nine highly polymorphic microsatellite primers for *Platycladus orientalis* (Cupressaceae). American Journal of Botany 99 (7). https://doi.org/10.3732/ajb.1100583

Supplementary materials

Suppl. material 1: Length distribution of assembly transcript doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: images
Brief description: Length distribution of assembly transcript for *C. konishii*.
Download file (50.27 kb)

Suppl. material 2: Length distribution of assembly unigenes doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Images
Brief description: Length distribution of assembly unigenes for *C. konishii*.
Download file (45.68 kb)

Suppl. material 3: Distribution of species search of unigenes doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: images
Brief description: Distribution of species search of unigenes against the Nr database of *C. konishii*.
Download file (96.42 kb)

Suppl. material 4: Distribution of different repeat type classes doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: images
Brief description: Distribution of different repeat type classes in *C. konishii* transcriptome.
Download file (18.25 kb)

Suppl. material 5: Mantel test of genetic distance and geographical distance doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: images
Brief description: Mantel test of genetic distance and geographical distance of three *C. konishii* populations.
Download file (29.28 kb)

Suppl. material 6: Table S1. Sampling location doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S1
Brief description: Table S1. Sampling location for *C. konishii* from Vietnam in the present study.
Download file (13.65 kb)

Suppl. material 7: Table S2. The main identified essential oil biosynthetic genes doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S2
Brief description: Table S2. The main identified essential oil biosynthetic genes from *C. konishii* unigenes.
Download file (14.23 kb)

Suppl. material 8: Table S3. Summary of analyses of expressed sequence doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S3
Brief description: Table S3. Summary of analyses of expressed sequence Tag–Simple Sequence repeat (EST-SSRs) in *C. konishii*.
Download file (13.72 kb)

Suppl. material 9: Table S4. Frequency distribution of SSRs doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S4
Brief description: Frequency distribution of SSRs, based on motif types in *C. konishii* transcriptome.
Download file (22.25 kb)

Suppl. material 10: Table S5. Primer sequences, repeat motif, size range of alleles and annealing temperature doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S5
Brief description: Table S5. Primer sequences, repeat motif, size range of alleles and annealing temperature (Tm) of 10 polymorphic EST-SSR markers developed for *C. konishii*.
Download file (15.35 kb)

Suppl. material 11: Table S6. Pairwise genetic differentiation (F_{ST}) between populations doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S6
Brief description: Table S6. Pairwise genetic differentiation (F_{ST}) between populations for *C. konishii* species.
Download file (13.57 kb)

Suppl. material 12: Table S7. Number of individuals for each population assigned doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S7
Brief description: Table S7. Number of individuals for each population assigned. Each cluster was obtained from DAPC without prior information.
Download file (13.30 kb)

Suppl. material 13: Table S8. Percentage of ancestry doi

Authors: Bich Hong Ha, Mai Phuong Pham, Quoc Khanh Nguyen, Thi Tuyet Xuan Bui, Dinh Giap Vu, Syed Noor Muhammad Shah, Dinh Duy Vu
Data type: Table S8
Brief description: Table S8. Percentage of ancestry for three *C. konishii* populations.
Download file (13.41 kb)